Project 4

West Nile Virus Prediction

GLADYS, RUBY, CHEE HOWE, JOHN

Agenda

- Problem Statement
- Methodology
- Data Cleaning
- EDA and Feature Engineering
- Model Selection and Evaluation
- Cost-Benefit Analysis
- Conclusion and Recommendations

Problem Statement

- Problem:
  - Due to surge of West Nile virus, the Chicago Department of Public Health is concerned about public health, as well as cost to the city.
- Aim:
  - Predict WNV presence in a location
  - Highlight conditions that correlates to WNV presence
  - Conduct a Cost-Benefit analysis to determine when and where to implement control measures

Methodology

DATA CLEANING AND FEATURE ENGINEERING

- Impute missing values
- Engineered new features including rolling windows, feature weights and localising features

MODEL SELECTION AND ANALYSIS

- Implemented regression, bagging and bossting models
- Used ROC_AUC score as a selection metric
- Identified important features

FINDINGS AND COST-BENEFIT ANALYSIS

- Studied weather, conditions and locations that correlate to WNV presence
- Weighed the medical cost against the cost of control measures

Data Cleaning

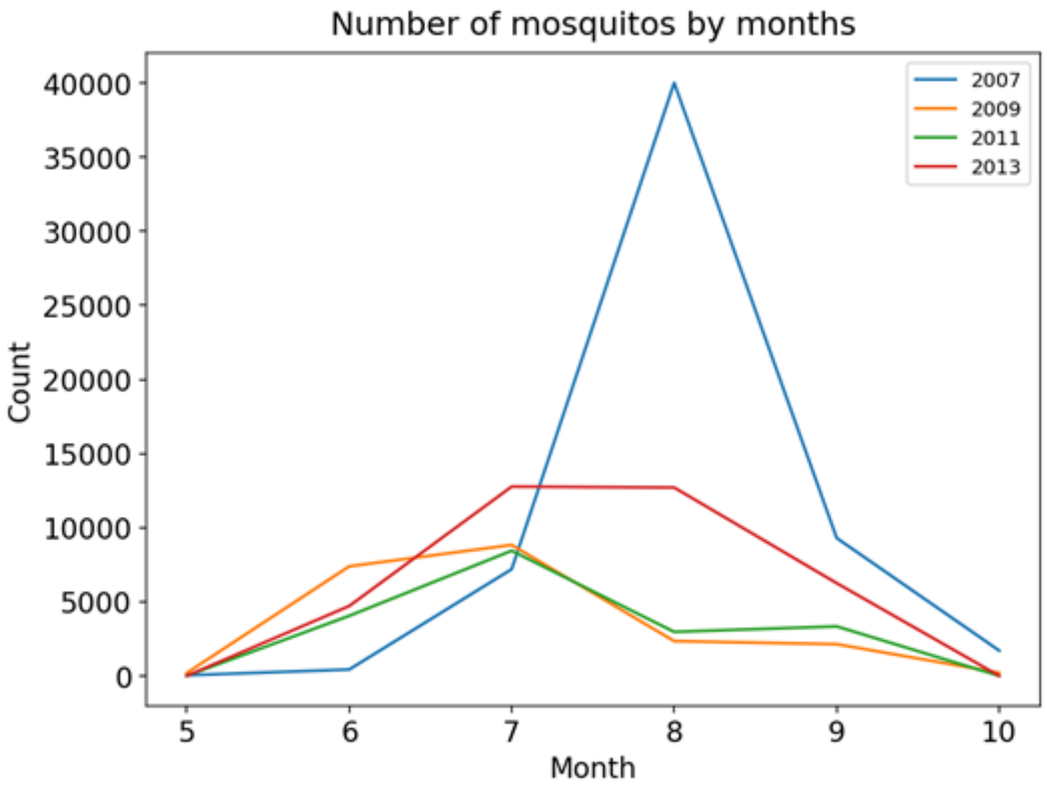
- Impute missing weather values by cross referencing both weather stations

- Utilise near-constant differences between features in each weather station.
  - Difference in sea level and station pressure is nearly constant at ~0.72 and ~0.65 for station 1 and 2 respectively

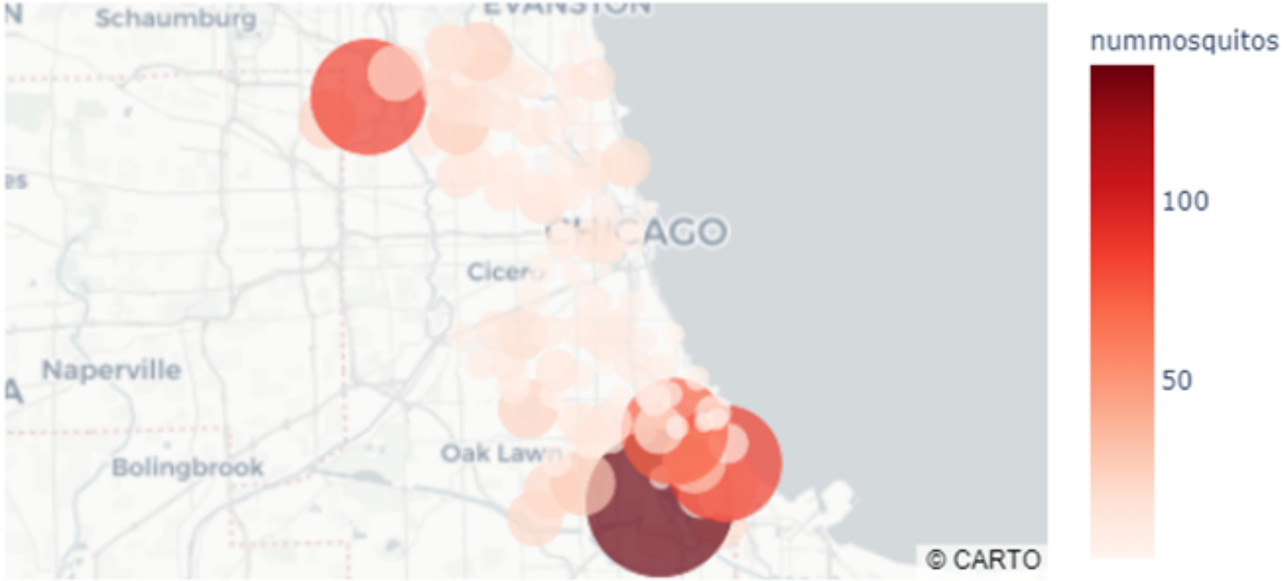
- Utilise rolling windows for the following features;
  - Precipitation
  - Temperature (average, min, max)
  - Dew point

EDA and Feature Engineering

EDA - Number of mosquitos

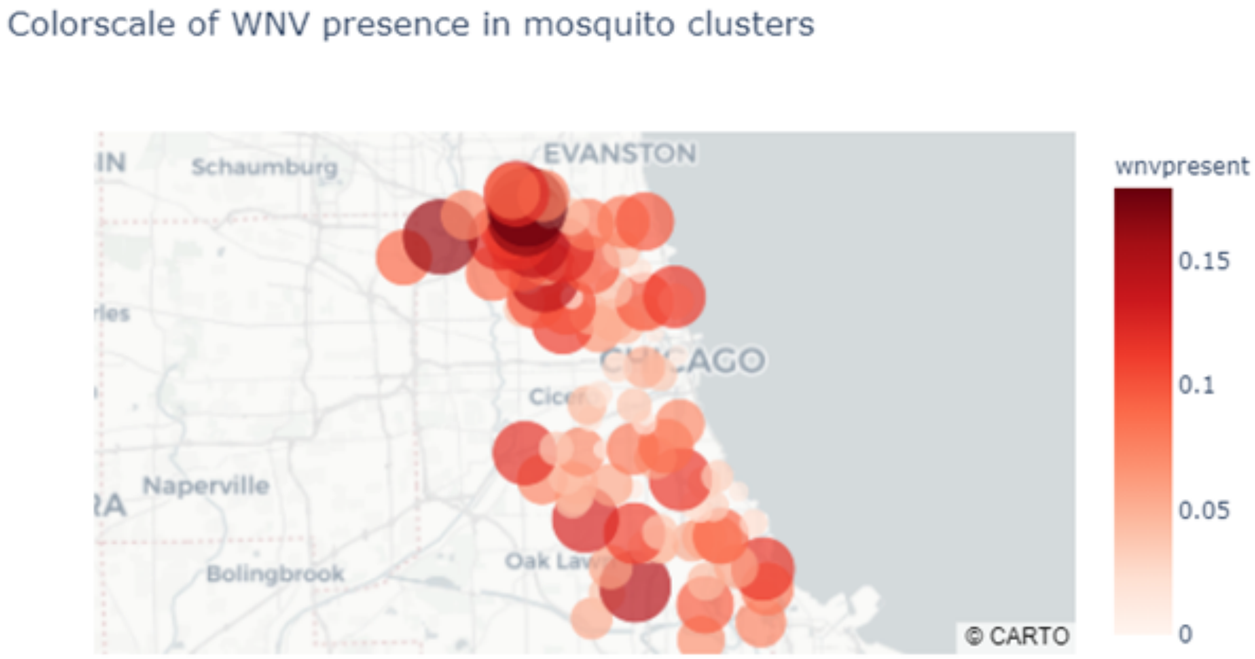


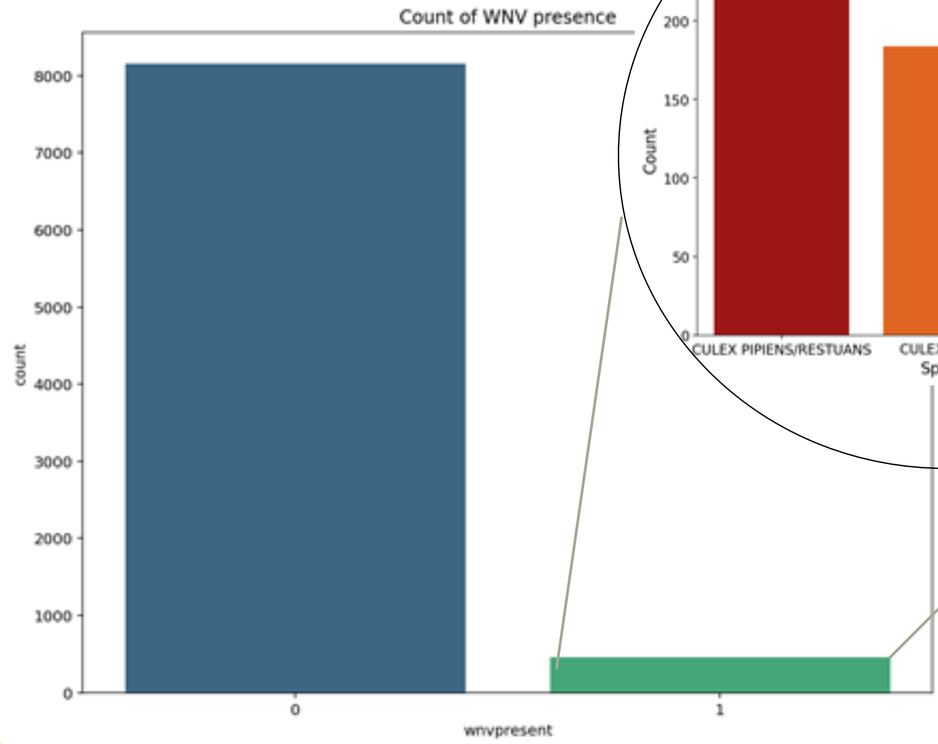
Locations with mosquito clusters

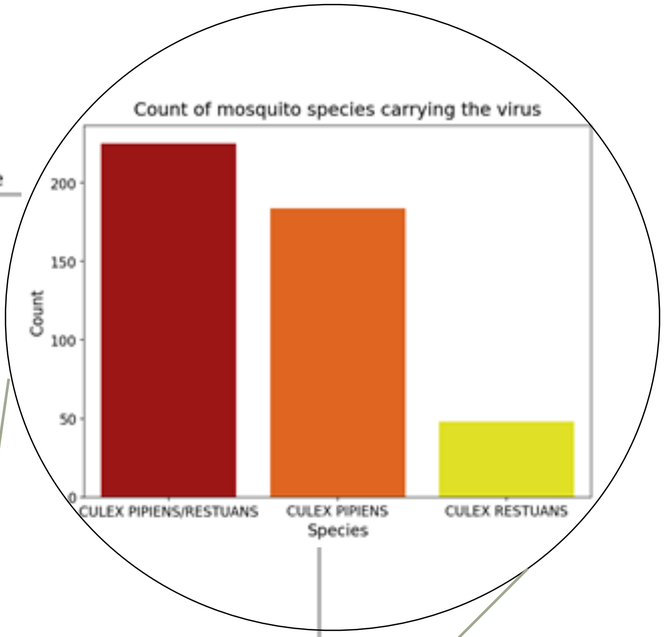


- MOSQUITO CLUSTERS IN TERMS OF SIZE. LARGER AND DARKER PLOTS SHOWS HEAVIER HIT AREAS IN AVERAGE MOSQUITO COUNT
- WE CAN IDENTIFY AT LEAST 4 HEAVILY HIT AREAS WITH HIGH NUMBER OF MOSQUITOS
- HEAVIEST HIT MONTHS OVER THE YEARS ARE JUNE TO SEP

EDA - Species

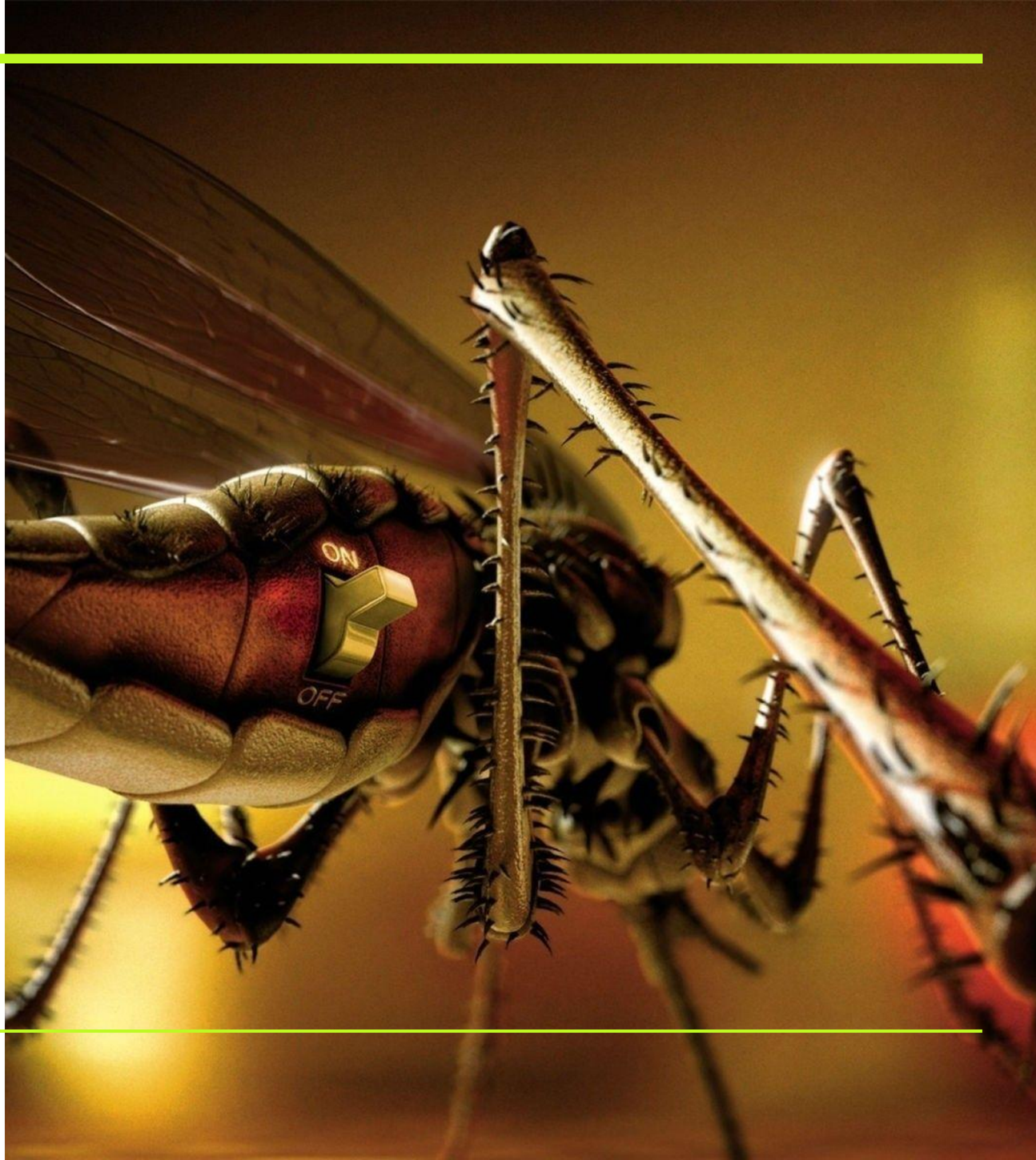




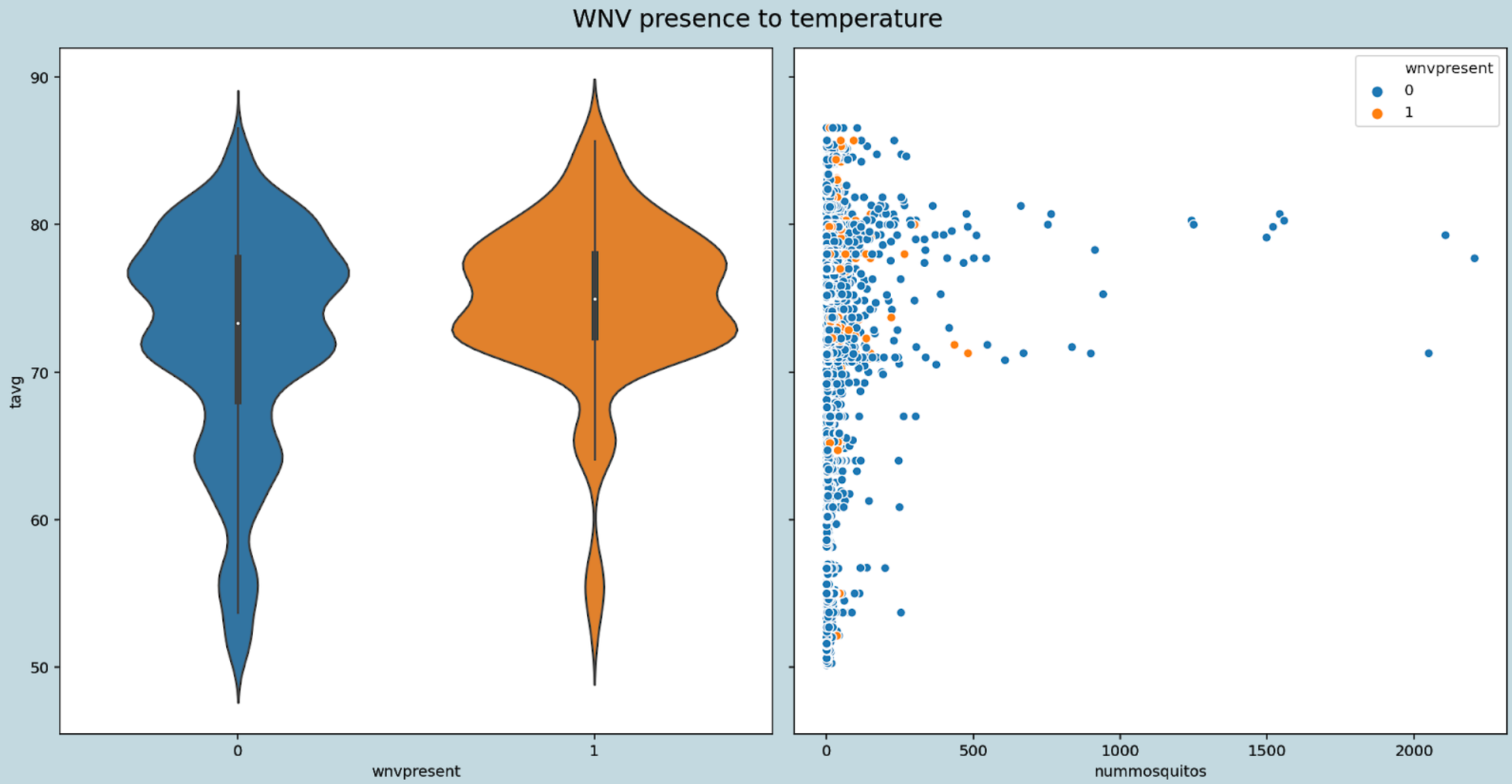


- RANKING OF CLUSTERS IN TERMS OF SIZE. LARGER AND DARKER PLOTS SHOWS HEAVIER HIT AREAS WITH AVERAGE WNV POSITIVE CASE
- IMBALANCED CLASSES OBSERVED
- NUMBER OF MOSQUITOS IS NOT THE MAIN CAUSE OF CONCERN FOR INCREASING NUMBER OF WNV
- 3 MAIN SPECIES FOUND TO BE WNV POSITIVE

Look at other conditions to find the link to West Nile virus



Relationship between Temperature and WNV presence



A high average temperature sees an increasing frequency of WNV positive mosquitos

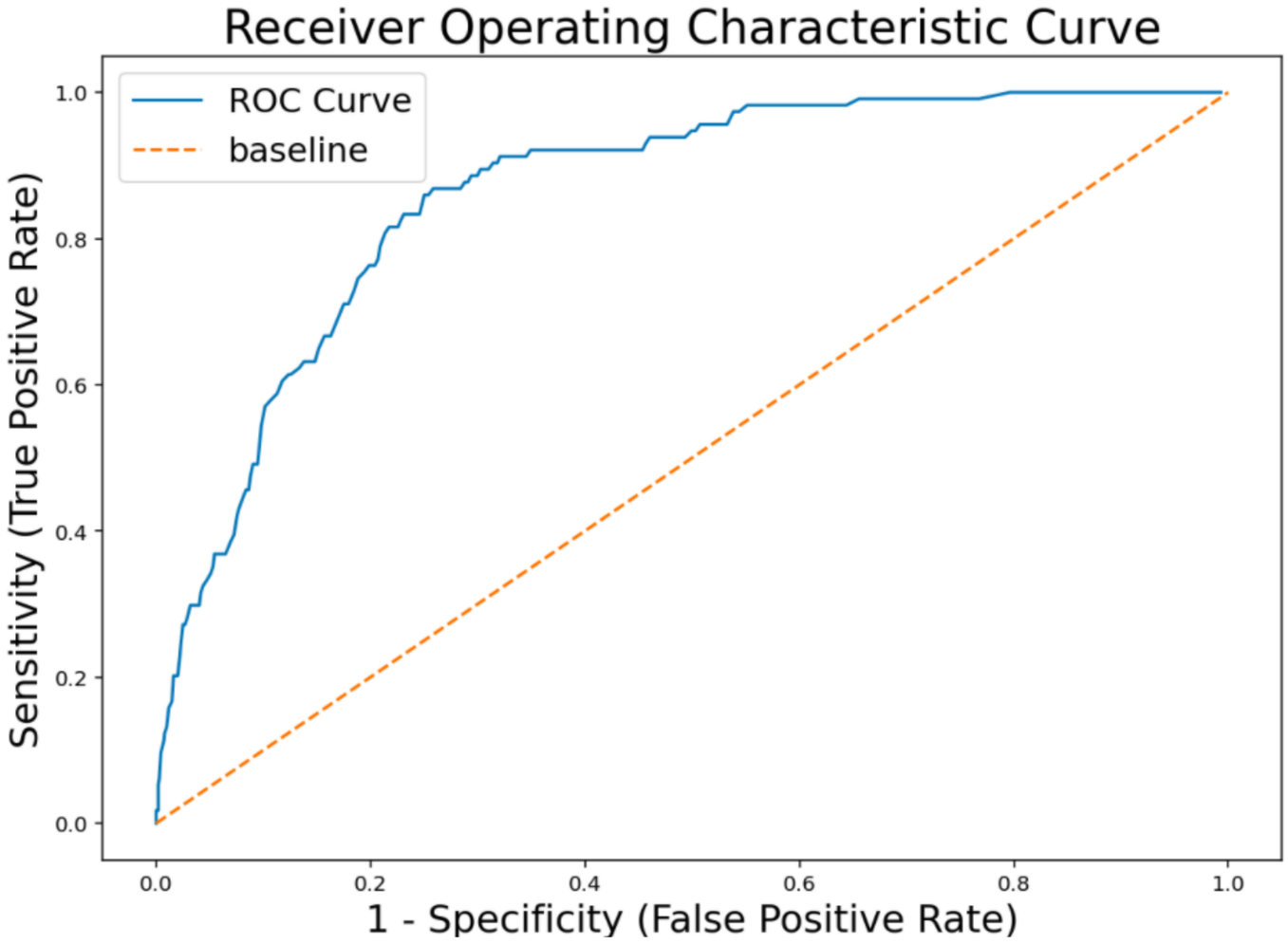
Our Feature Engineering Approach

- Ratio of each weather station from each trap to calculate the localised weather condition
- Find the probability of WNV positive in each of species
- Find the proportion of number of mosquitos in each species
- Get the harmonic mean across both to engineer the final weight allocated to each species
- Allocate weights in terms of WNV occurrence in each of the observed months
- Get the distance away from the top 4 hotspots of WNV cases and given each location new features base on that distance

Model Selection and Evaluation

MODEL	ROC-AUC score	Recall score	False Negatives
Logistic Regression (Our Baseline)	0.818		
Random Forest	0.837	0.833	19
XGBoost	0.834	0.728	31

Model Evaluation

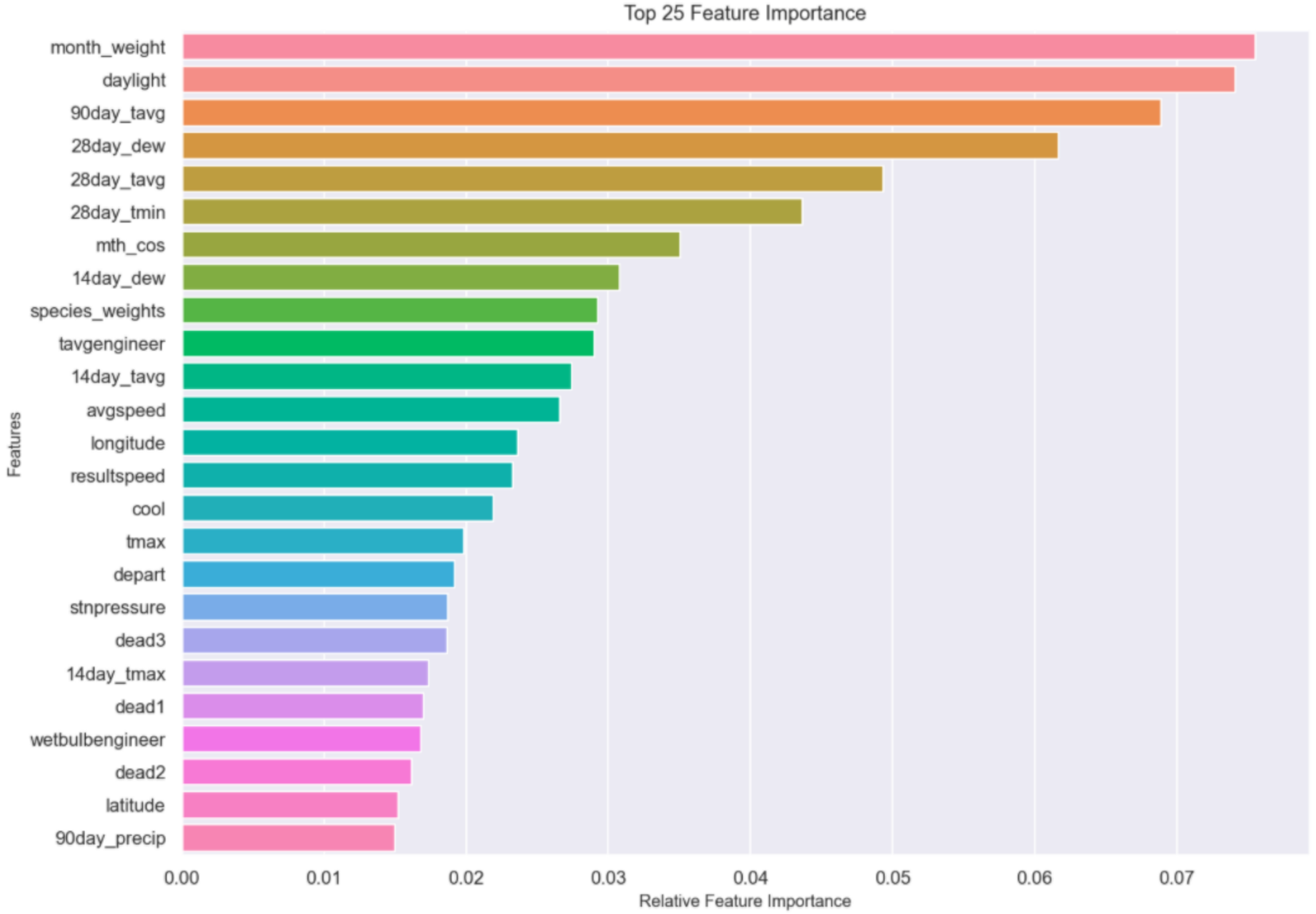


RELATIVELY HIGH ROC-AUC OF 83.7%

Confusion Matrix

		True Label	
		Negative WNV	**Positive WNV**
Predicted Label	**Negative WNV**	1558	19
	Positive WNV	481	95

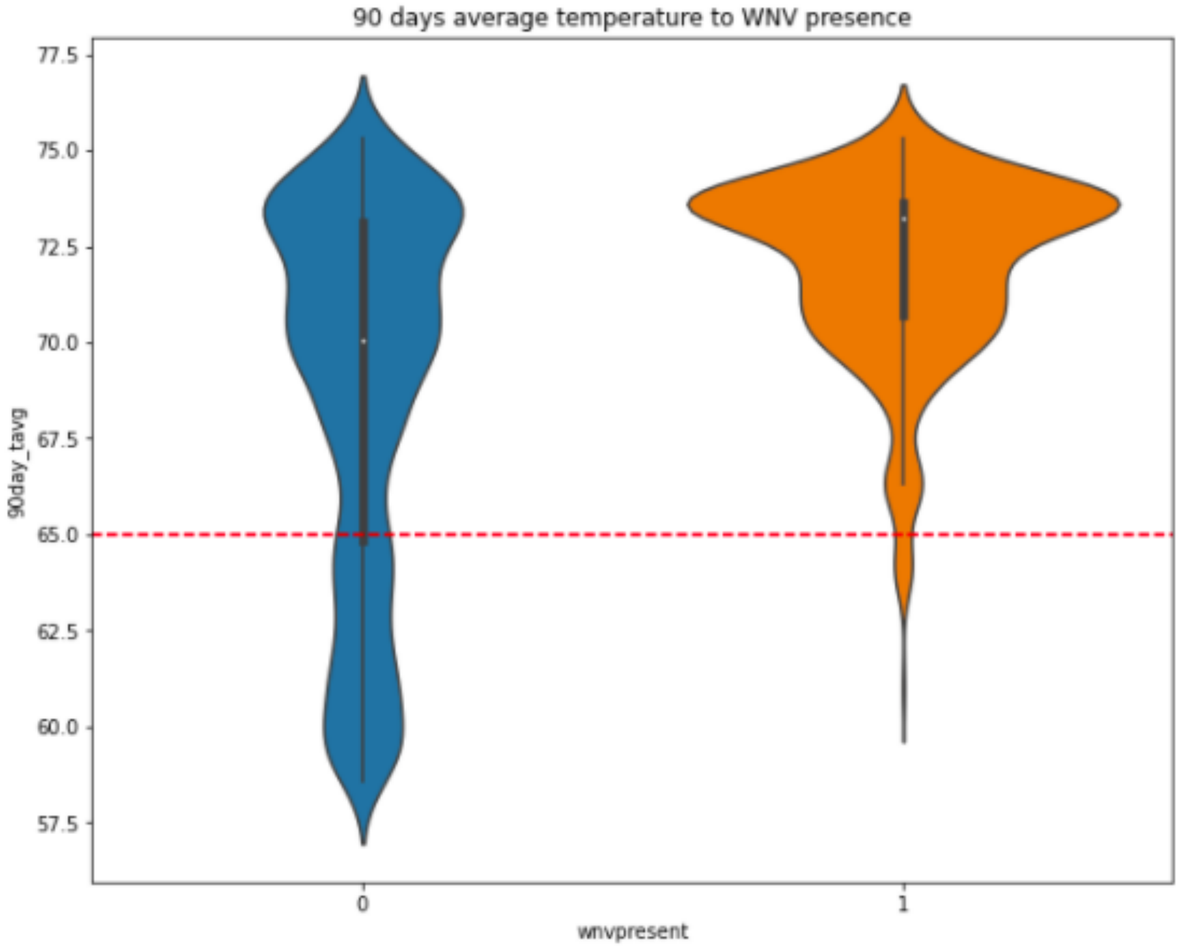
RELATIVELY HIGH RECALL RATE OF 83.3% , LOW NUMBER OF FALSE NEGATIVES

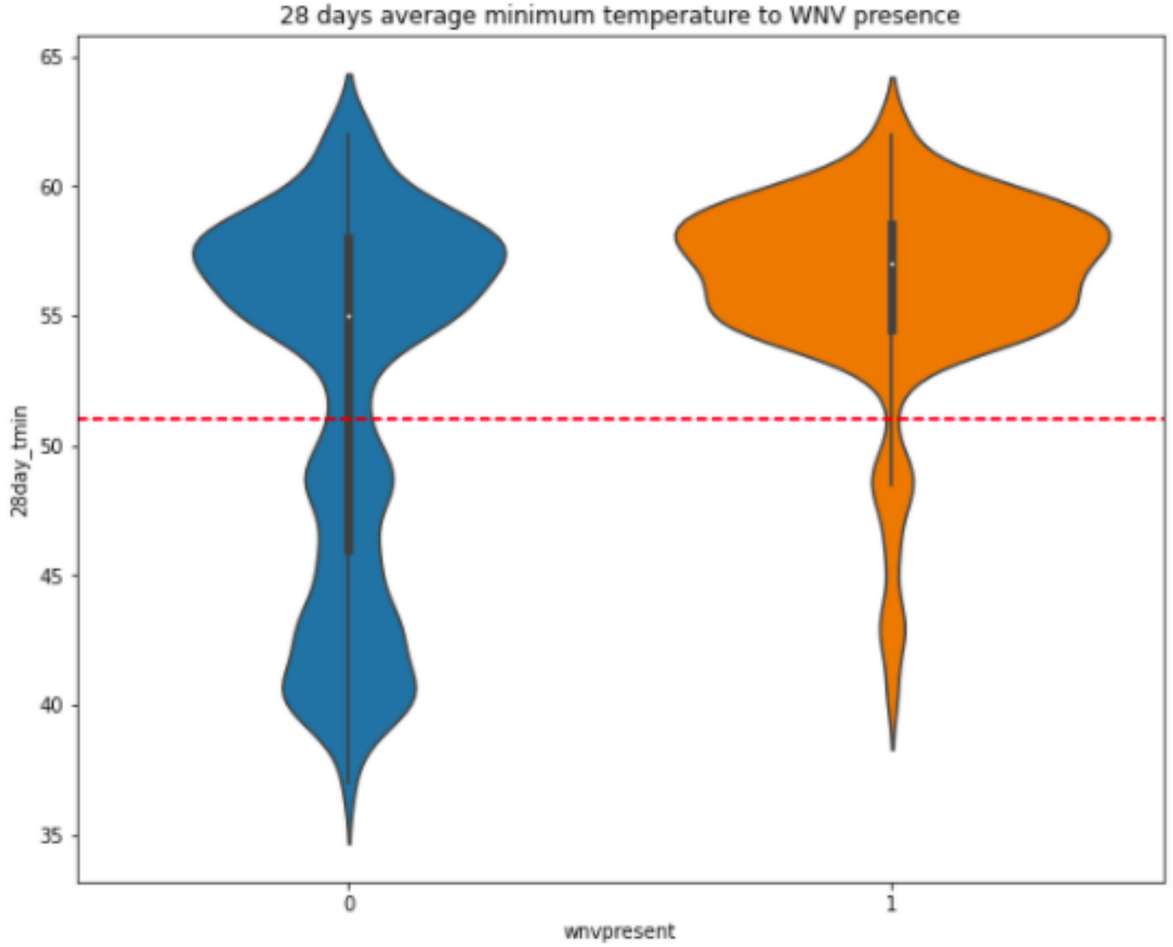


Model Interpretation

- Highly-ranked important features:
  - Month
  - Temperature (minimum, average)
  - Daylight
  - Distance to WNV hotspots

Findings


90 days average temperature to WNV presence
77.5
75.0
72.5
70.0
67.5
65.0
62.5
60.0
57.5
90day_tavg
0
1
wnvpresent


28 days average minimum temperature to WNV presence
65
60
55
50
45
40
35
28day_tmin
0
1
wnvpresent

Cost-Benefit Analysis

Cost-Benefit Analysis

1. WHEN TO SPRAY?

- Month
- Frequency

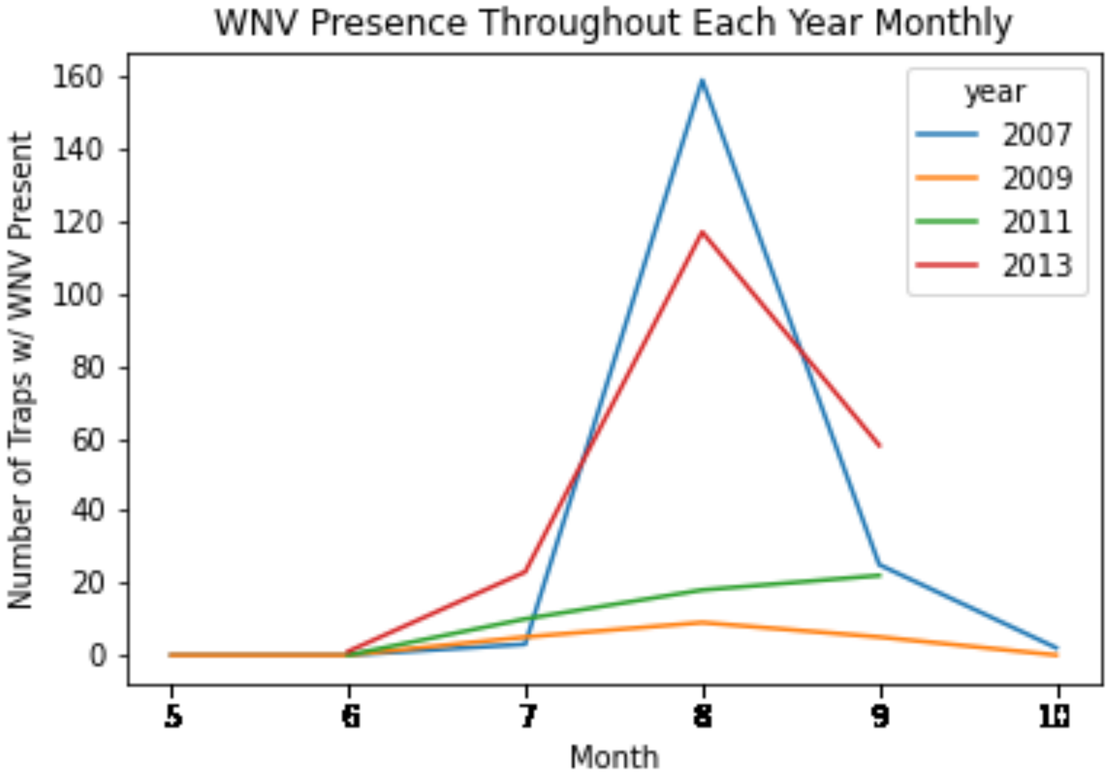
2. WHERE TO SPRAY?

- Area to be sprayed
- Various coverage levels

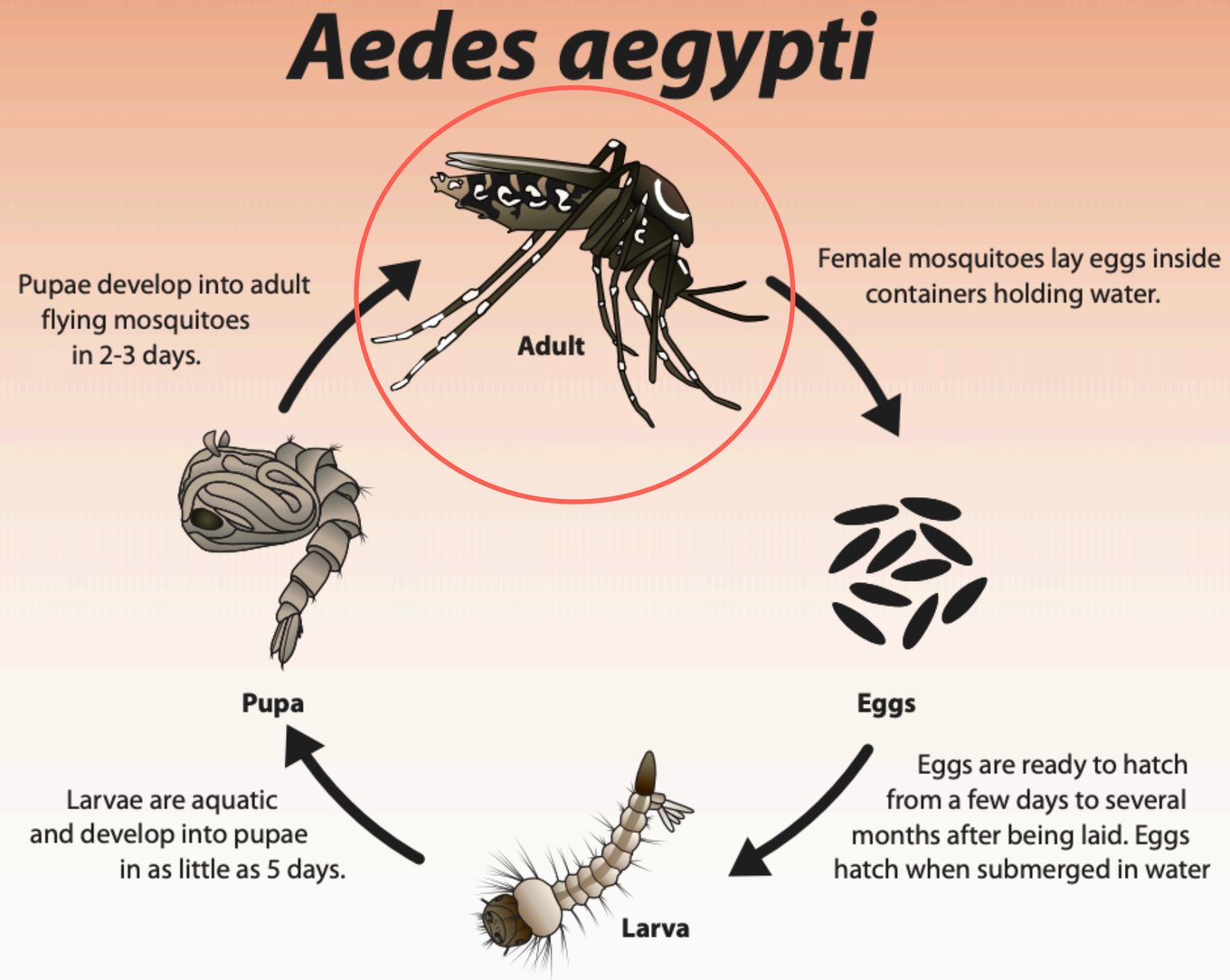
3. HOW MUCH?

- Spray cost
- Potential benefit

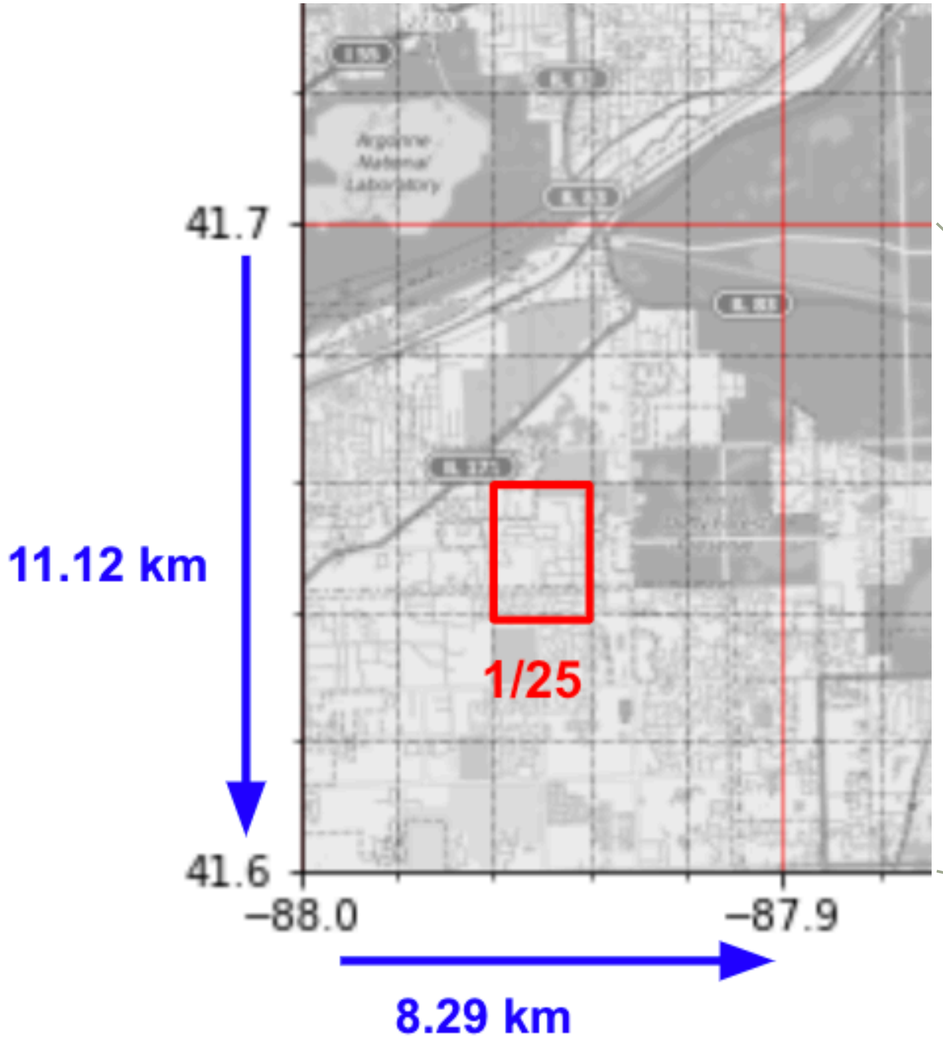
1. When to spray?

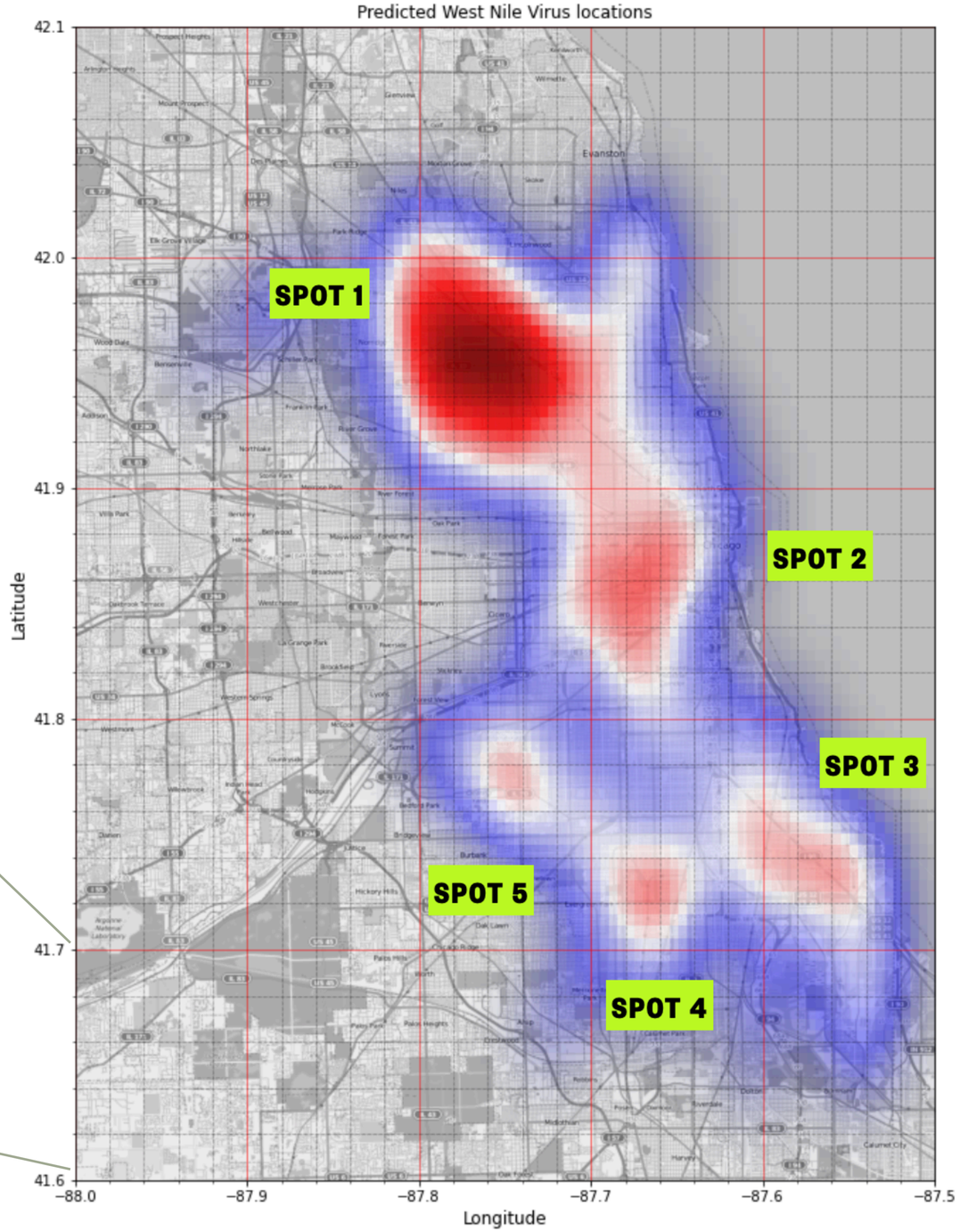


- **MONTHS: JUL - SEP**
- **FREQUENCY: ONCE A WEEK**



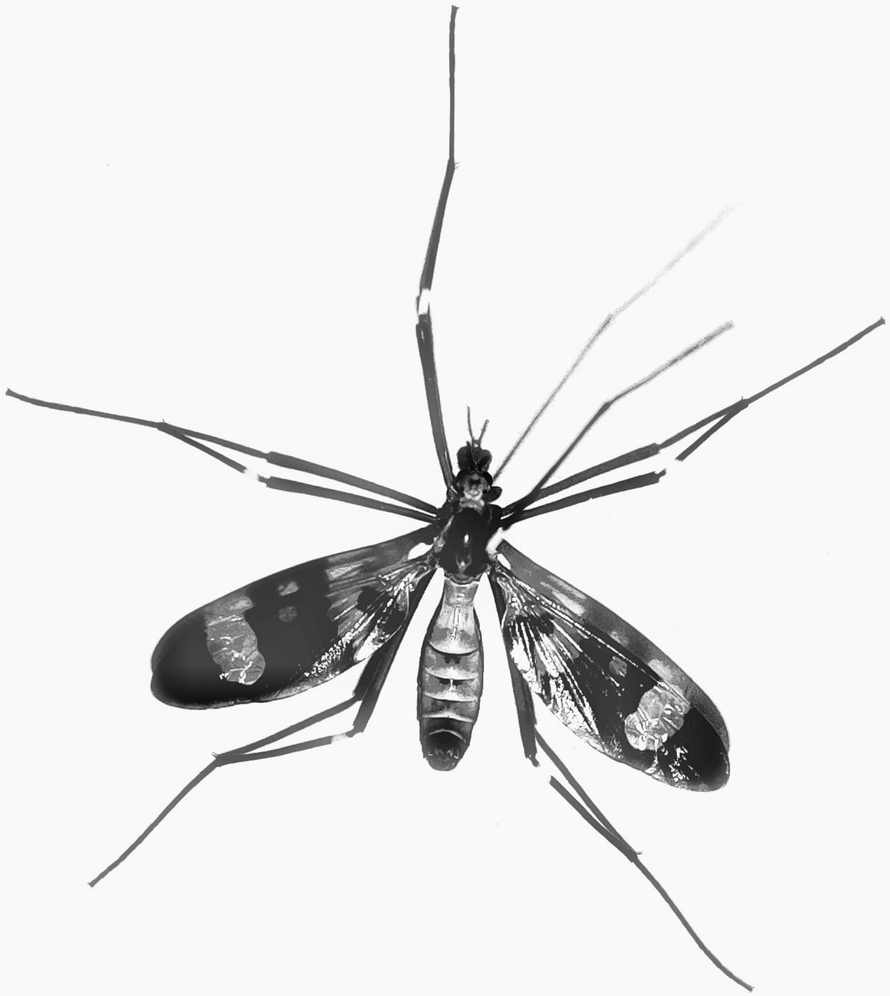
2. Where to spray?





3. How much?

- Vector control cost
- ≈ USD 701,790
- Per spray
- Per km
- ≈ USD 245.2



3. How much?

PLAN	COVERAGE	AREA TO BE SPRAYED	COST PER SPRAY PER KM²	TIMES OF SPRAYING	TOTAL SPRAY COST
Plan 1	Spot 1	92 km²	USD 245.2	12	USD 270,906
Plan 2	Spot 1 + 2	129 km²			USD 378,981
Plan 3	All 5 spots	188 km²			USD 552,229

Coverage level

PLAN 1

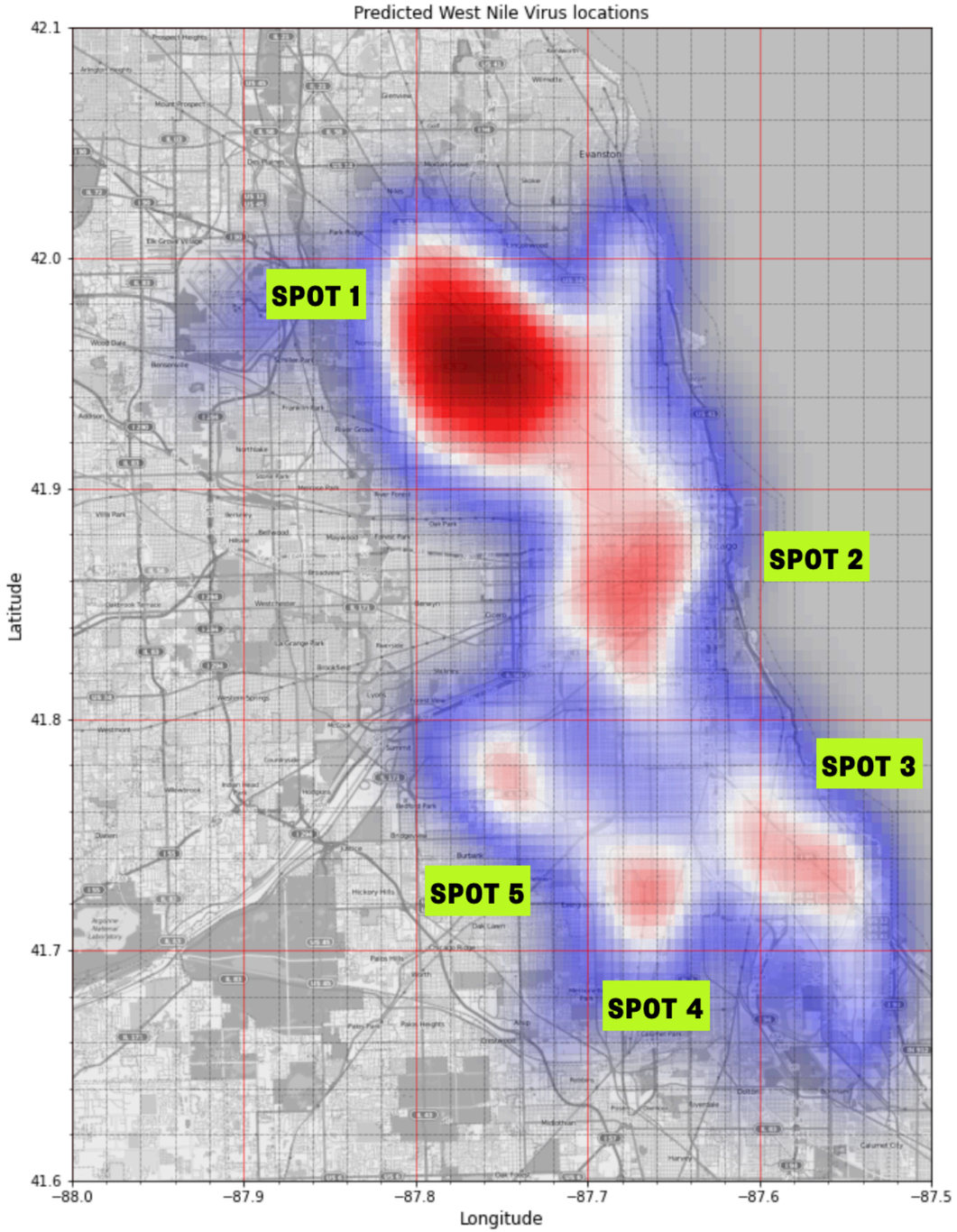
SPOT 1
USD 270,906

PLAN 2

SPOT 1 + 2
USD 378,981

PLAN 3

ALL 5 SPOTS
USD 552,229



YEAR	WNV CASE	COVERAGE	AVOIDED LOSS	SPRAY COST	POTENTIAL BENEFIT
2010	47	Plan 3 All	USD 858,972	USD 552,229	USD 270,906
2011	24	Plan 2 / ?	USD 438,624	USD 378,981	USD 378,981
2012	229	Plan 3 All	USD 4,185,204	USD 552,229	USD 3,632,975
2013	66		USD 1,206,216		USD 653,987
2014	31		USD 566,556		USD 14,327
2015	36		USD 657,936		USD 105,707
2016	108		USD 1,973,808		USD 1,421,579

30 cases

To outweigh spray cost

3.6 million

2012: 229 cases

Conclusion and Recommendations

Conclusion

1. FEATURES:
   - WEATHER
   - LOCATION
   - SEASON
   - POPULATION DENSITY

2. CBA

Recommendations

1. MOSQUITO CONTROL
   - COVERAGE
   - SUMMER SEASON
2. SURVEILLANCE: SPECIES
3. PUBLIC EDUCATION